Received: 3 January 2023 | Revised: 17 April 2023 | Accepted: 5 May 2023

DOI: 10.1002/psp4.12998

ARTICLE

A systematic approach towards missing lab data in electronic health records: A case study in non-small cell lung cancer and multiple myeloma

Arjun Sondhi[1] | **Janick Weberpals**[2] | **Prakirthi Yerram**[1] | **Chengsheng Jiang**[1] | **Michael Taylor**[3] | **Meghna Samant**[1] | **Sarah Cherng**[1]

[1]Flatiron Health, Inc., New York, New York, USA

[2]Hoffmann-La Roche, Basel, Switzerland

[3]Genentech, San Francisco, California, USA

Correspondence
Arjun Sondhi, Flatiron Health, 233 Spring St, 5th Fl, New York, NY 10013, USA.
Email: arjun.sondhi@flatiron.com

Abstract

Real-world data derived from electronic health records often exhibit high levels of missingness in variables, such as laboratory results, presenting a challenge for statistical analyses. We developed a systematic workflow for gathering evidence of different missingness mechanisms and performing subsequent statistical analyses. We quantify evidence for missing completely at random (MCAR) or missing at random (MAR), mechanisms using Hotelling's multivariate *t*-test, and random forest classifiers, respectively. We further illustrate how to apply sensitivity analyses using the not at random fully conditional specification procedure to examine changes in parameter estimates under missing not at random (MNAR) mechanisms. In simulation studies, we validated these diagnostics and compared analytic bias under different mechanisms. To demonstrate the application of this workflow, we applied it to two exemplary case studies with an advanced non-small cell lung cancer and a multiple myeloma cohort derived from a real-world oncology database. Here, we found strong evidence against MCAR, and some evidence of MAR, implying that imputation approaches that attempt to predict missing values by fitting a model to observed data may be suitable for use. Sensitivity analyses did not suggest meaningful departures of our analytic results under potential MNAR mechanisms; these results were also in line with results reported in clinical trials.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

Missing data is a common issue in real-world data (RWD), and appropriate analytic methods depend on the underlying missingness mechanism, which is untestable. Methods are generally applied in an ad hoc manner, and do not convey how robust results may be to departures from assumptions.

Arjun Sondhi and Janick Weberpals contributed equally to this work.

WHAT QUESTION DID THIS STUDY ADDRESS?
This study addresses how to implement diagnostic and analytic methods for missing data in RWD.
WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?
This study demonstrates how to gather evidence for different missingness mechanisms, which then informs subsequent data analysis methods. In particular, this study recommends a systemic approach, with an emphasis on sensitivity analyses to demonstrate the robustness of scientific conclusions. We also differentiate between two distinct mechanisms for missing not at random.
HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?
This study serves as a guide for handling missing data, which is common in RWD analyses. Our workflow allows analysts to connect evidence for missingness approaches to analytic methods, providing transparency in the robustness of reported results.

INTRODUCTION

Healthcare data collected outside of clinical trials (real-world data [RWD]) are increasingly used in research and clinical drug development, including regulatory decision making.[1] Electronic health records (EHRs), in particular, are an important backbone for RWD research as they provide detailed clinical information on a patient's disease journey. In cancer research, it was found that under routine care settings, certain laboratory tests show a high prognostic value across different tumor types.[2] Ideally, complete and perfectly measured laboratory data would provide the means to implement inclusion/exclusion criteria for selecting real-world cohorts of patients, define end points and covariates for assessing treatment effectiveness or adverse events, and support the development of prognostic scores. However, laboratory values in EHR-derived data are often subject to a high degree of missingness due to varying data capture and screening decisions in routine clinical care, posing a challenge for the use of RWD in statistical analyses.

When analyzing data subject to missingness, approaches based on complete case data analyses may end up being biased if data missingness does not occur completely at random, but is associated with underlying features. In this case, imputation approaches where missing values are predicted using additional variables collected in the dataset, may allow for unbiased analytic results with less uncertainty, due to using the entire dataset.[3] However, if the missing values are systematically different from the observed ones in a way that cannot be appropriately modeled, then imputation can lead to biased results. Sensitivity analyses can model departures from ignorable missingness, and the appropriate analysis strategy depends on the missingness mechanism present in the data.

In this paper, motivated by a desire to mitigate the impact of missingness in laboratory data derived from EHRs, we outline a systematic workflow for characterizing missing data mechanisms, and performing subsequent statistical analyses. In general, it is not possible to formally test data for different missingness mechanisms because these inherently depend on the distribution of the unobserved values. Instead, we propose the use of diagnostics that provide evidence for certain mechanisms over others which may be used alongside specific domain knowledge of the underlying data generating mechanisms. First, we outline our basic assumptions, which we validate in a simulation study. We then apply our approach to analyzing two exemplary case studies from an EHR-derived oncology database, illustrating how to analyze RWD subject to missingness and interpret subsequent results.

METHODS

Background

For this study, we consider an EHR-derived dataset containing a single laboratory variable subject to missingness and other variables that are fully observed. However, our methods can be extended to the multivariate missingness setting. We denote the true laboratory variable as lab, the observed laboratory value with missingness as $\mathsf{lab}_{\mathrm{obs}}$, and the remaining variables as X. These variables may include covariates and outcomes. Additionally, we define M_{lab} as a

binary missingness indicator for lab; $M_{\text{lab}} = 1$ if lab is missing, and 0 otherwise.

We first define missing completely at random (MCAR). Under this missingness mechanism, the probability of a patient having a missing laboratory value does not depend on their true laboratory value or on any other (observed or unobserved) variables. For example, if laboratory values measured on a certain day were deleted due to a technical failure, this would result in the data being MCAR. Given MCAR data, a complete case analysis will generally provide unbiased results, and only suffer from a loss of precision due to having a smaller sample size.[4] In such cases where sample size and the associated loss in statistical efficiency are a concern, multiple imputation can be beneficial even under MCAR if the missingness is mostly observed in exposure or confounders or if enough auxiliary variables (i.e., variables that are not part of the primary analysis but show correlations to the partially observed variable), are available.[5] As a result, the retention of incomplete cases will increase statistical efficiency and a multiple imputation approach typically provides more realistic estimates of the variance as it also accounts for the extra variance caused by the missing data (between-imputation variance).[3]

Under the mechanism of missing at random (MAR), the probability of a patient having a missing laboratory test does not depend on their true laboratory value, but does depend on observed variables $\mathbf{X}$ (e.g., age, sex, therapy type, and outcomes). An example of MAR laboratory values would be if older patients were more likely to receive tests, and age was measured in the dataset. MAR data are usually assumed in order to obtain unbiased analytic results under imputation procedures. Generally, the performance of multiple imputation procedures, however, additionally depends also on other factors, like the variable type that is partially observed (exposure, outcome, or confounders)[5] and the presence of auxiliary variables.[6]

Data are considered to be missing not at random (MNAR) in the absence of MCAR or MAR.[7] We consider two different types of relationships yielding MNAR data. First, if the probability of a patient having a missing laboratory test does not depend on their true laboratory value, but does depend on some unobserved variables U, this would be considered MNAR. This can arise, for example, if patients with a certain biomarker not observed in the dataset are more likely to have missing laboratory values. Second, data are also considered MNAR if the missingness probability is directly affected by the true laboratory values; for example, if patients with a history of normal laboratory values are systematically less likely to be tested. These two MNAR mechanisms have different implications for imputation accuracy and the resulting analytic bias. We refer to the former mechanism as MNAR-unobserved and the latter as MNAR-value.

For a graphical overview of all mechanisms, see Figure 1.

Proposed workflow and statistical analysis

Depending on the above described missingness mechanisms, different analytical methods may be used to address missingness in statistical analysis. However, selecting the appropriate approach requires sufficient evidence for a particular missingness mechanism. To gather evidence for the different assumptions, we propose the following workflow (Figure 2).

Step 1: Assess evidence for MCAR

Under MCAR, patient characteristics between patients with and without an observed laboratory test should be balanced as the missingness is independent of both the observed and unobserved data. On a variable-by-variable basis this can be examined by comparing the standardized mean differences (SMDs) between patients with and without an observed laboratory test; an absolute standardized difference less than 0.1 is conventionally considered to indicate balance and would give evidence for the missingness being MCAR.[8] Further, Little's chi-squared test,[9] which takes into account possible patterns of missingness across all variables in the dataset, can be applied. Rejection of the null hypothesis of this test would provide sufficient evidence to indicate that the data are (globally) not MCAR and one would continue with step 2 (examining MAR). In case all other variables are fully observed and one is only interested in examining the missingness of the lab itself, Little's test is equivalent to a Hotelling's multivariate t-test,[10] which examines variable differences conditional on having an observed laboratory test or not. As the power of statistical hypothesis tests can be influenced by sample size, the combined investigation along with SMDs is recommended. Given a lack of evidence against MCAR, both complete case analysis and multiple imputation approaches may be feasible approaches that result in unbiased estimates for further downstream analyses.[7,11]

Step 2: Assess evidence for MAR

Given a lack of evidence for MCAR, the next possible hypothesis would examine an MAR scenario where the

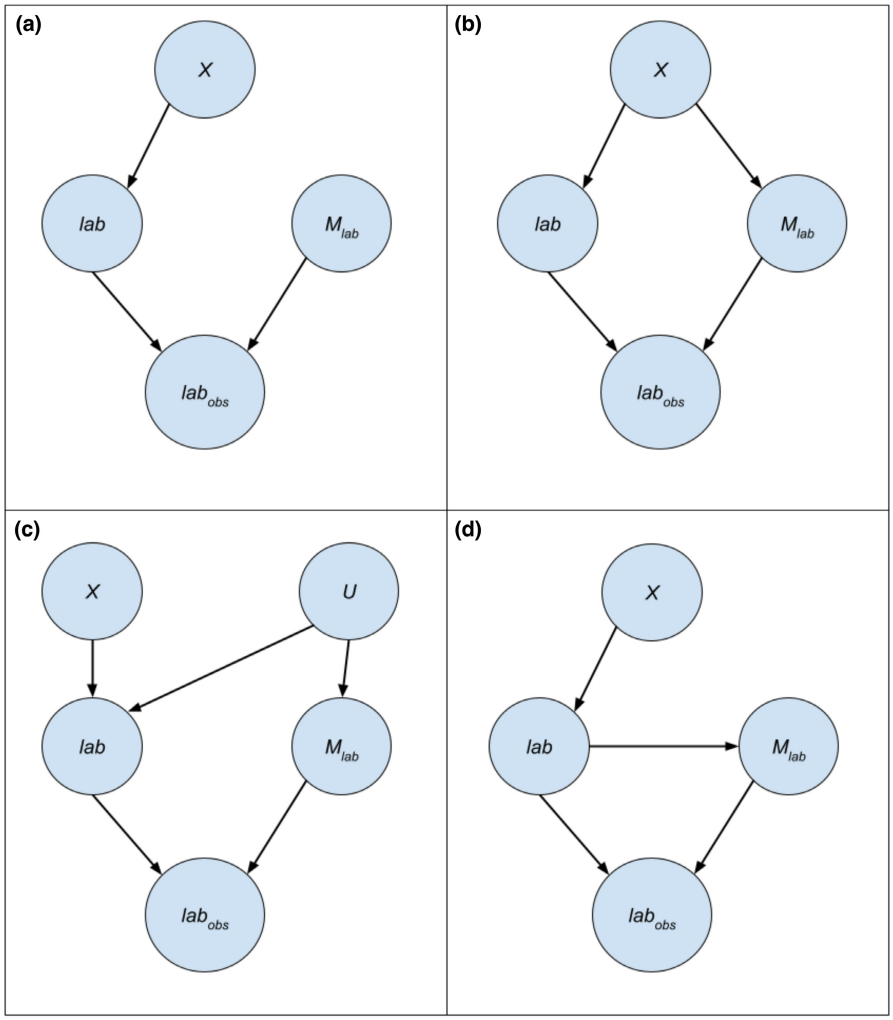

FIGURE 1 M-graph[34] representations of the four missingness mechanisms considered, with lab_{obs} representing observed laboratory values subject to missingness. (a) MCAR: Missingness (M_{lab}) is independent of the true laboratory values (lab) and any other variables X. (b) MAR: Missingness (M_{lab}) depends on observed variables X but is independent of true laboratory values (lab). (c) MNAR-unmeasured: Missingness (M_{lab}) depends on unobserved variables U but is independent of true laboratory values (lab). (d) MNAR-value: Missingness (M_{lab}) depends on true laboratory values (lab). A directed edge represents a causal effect of a variable on another. MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

missingness could be predicted by using a classification model using observed variables X. As part of this workflow, we will use a random forest (RF) classification model due to its ability to implicitly model nonlinear and non-additive relationships between observed variables.[12] The performance of the classification model to predict the missingness is measured by the area under the receiver operating characteristic curve (AUC). A sufficiently high AUC with values meaningfully greater than 0.5 would give evidence that the missingness is dependent on the largely observed variables and hence would give some indication for the missingness being MAR. Given a relationship between missingness and observed data, multiple imputation would be an appropriate approach to handle missingness, whereas a complete case analysis would likely result in biased estimates.

Step 3: Apply not at random fully conditional specification sensitivity analysis to assess robustness to MNAR mechanisms

If MNAR is suspected as a plausible missingness mechanism, we recommend sensitivity analyses of the final statistical model using the not at random fully conditional specification (NARFCS) algorithm, as proposed by Tompsett.[13] In practice, because it is not possible to differentiate MNAR using observed data, applying NARFCS is useful to demonstrate the robustness of imputation-based approaches, even with evidence of MAR. The NARFCS procedure consists of shifting the imputations drawn at each iteration of a multiple imputation model under an MAR scenario by a user-specified quantity (sensitivity parameter δ) to reflect systematic departures of the missing

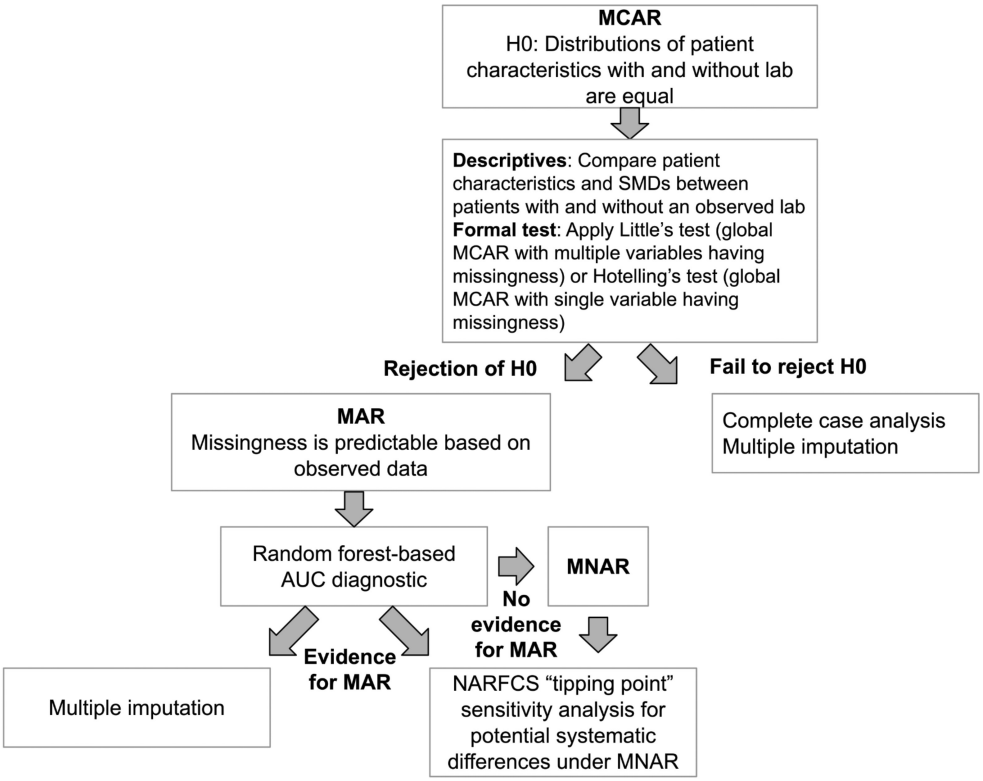

FIGURE 2 Illustration of systematic workflow to diagnose potential missingness mechanisms. AUC, area under the receiver operating characteristic curve; H0, null hypothesis; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random; NARFCS, not at random fully conditional specification; SMD, standardized mean difference.

data from the observed data distribution. It can be used as a "tipping point" sensitivity analysis to investigate how large a (plausible) systematic departure of missing data versus observed data would need to be to qualitatively change the conclusion of an analysis. The sensitivity parameter can be interpreted as the average difference between imputed and observed values of a variable, conditional on all remaining variables of the data. That is, for a single covariate, a shift in the sensitivity parameter may be interpreted as a shift to the intercept on the scale of the unit of the covariate. Given an analysis with sensitivity parameter $\delta = 0$, this would be equivalent to an ordinary imputation based on a fully conditional specification[13] without any sensitivity parameter.

This systematic workflow aims to provide evidence for certain mechanisms over others but the exact missingness mechanism is formally not identifiable from incomplete data.[4] This systematic workflow aims to provide evidence for certain mechanisms over others but due to several factors, such as the almost always unknown data generating mechanisms and the "unknown unknowns," the exact missingness mechanism is formally not identifiable.

Data sources and study design

Simulation design

This first part of the study used de novo simulated datasets to assess and validate the general assumptions of the proposed workflow. We simulated data corresponding to a comparative effectiveness study, according to the following steps.

1. Continuous covariates were generated as $Z \sim N(0, 1)$.
2. Binary values indicating if a laboratory measurement was within a normal range were generated as $P(\text{lab} = 1) = 0.9\, I(Z < 0) + 0.1\, I(Z > 0)$.
3. Binary treatment indicators were generated as $\text{logit}\, P(\text{trt} = 1) = \log(0.7)\, Z + \log(1.3)\, \text{lab}$.
4. Finally, survival and censoring times were generated independently from an exponential distribution with hazard function $\lambda = 0.3 \exp(\log(2)\, Z + \log(0.7)\text{lab} + \log(0.8)\, \text{trt})$.
5. The observed time to event time was then selected as the minimum of survival and censoring times.

We then simulated missingness in the laboratory variable according to five missingness mechanisms, induced as follows:

1. MCAR: Each laboratory value has a constant missingness probability of 0.5.
2. MAR v1: The probability of missingness depends on the covariate Z, that is, $P(M_{\text{lab}} = 1) = 0.9\, I(Z < 0) + 0.1\, I(Z \geq 0)$.
3. MAR v2: The probability of missingness depends on all observed variables, that is, $\text{logit}\, P(M_{\text{lab}} = 1) = 1 + 2\, Z + \log(0.5)\, \text{time} + \log(1.7)\, \text{trt}$
4. MNAR-unmeasured: The probability of missingness depends on all observed variables and an unobserved variable U, that is, $\text{logit}\, P(M_{\text{lab}} = 1) = 1 + 2\, Z + \log(0.5)\, \text{time} + \log(1.7)\, \text{trt} + 2U$. Here, we generate $U \sim N(0, 1)$ and laboratory tests as $\text{logit}\, P(\text{lab} = 1) = 3\, Z + 3\, U$.
5. MNAR-value: The probability of missingness depends on all observed variables and the true lab, that is, $\text{logit}\, P(M_{\text{lab}} = 1) = 2\, Z + \log(0.5)\, \text{time} + \log(1.7)\, \text{trt} + 2\, \text{lab}$.

Parameters were set to result in strong associations between missingness and other variables; this was done in order to induce a meaningful amount of analytic bias. For each simulation iteration, we created simulated datasets with missing data by setting laboratory values to be missing according to their assigned probability. Using these simulated datasets, we assessed the performance of Little's test and the AUC MAR diagnostic under the different missingness mechanisms over 200 simulation iterations for each mechanism. For Little's test, we reported the proportion of p values less than 0.05 for each missingness mechanism. In order to implement the AUC diagnostic, we fit a random forest model to each simulated dataset and tuned hyperparameters using five-fold cross-validation to maximize AUC. We summarized the results by averaging the AUCs for each missingness mechanism.

For each missingness mechanism, we also examined the estimation bias of fitting Cox regression models to the datasets subject to missingness, under complete case analysis and multiple imputation. Here, we treated estimates from Cox models fit to the complete datasets without missingness as the ground truth. For each covariate, we reported the distributions of absolute bias and confidence interval coverage probabilities of estimated hazard ratios. We also reported the L1 bias, defined as the sum of absolute biases across all hazard ratios. Multiple imputation was implemented using the mice package in R,[14] with logistic regression as the imputation method, and 100 multiple imputations performed.

Application of workflow to real-world cohorts

For the second part of this study, we derived two real-world cohorts from the nationwide Flatiron Health EHR-derived de-identified database, a longitudinal database, comprising de-identified patient-level structured (e.g., laboratory values and prescribed drugs) and unstructured data from ~280 US cancer clinics (~800 sites of care) curated via technology-enabled abstraction.[15,16] The majority of patients in the database originate from community oncology settings. Institutional review board approval of the study protocol was obtained prior to study conduct, and included a waiver of informed consent.

To illustrate the application of our workflow and the resulting impact of missingness of hemoglobin (Hgb) as an important prognostic laboratory test for real-world overall survival (rwOS) across several tumor types,[2,17,18] the following case studies on two exemplary exposure-outcome associations were performed. First, we selected a real-world cohort of patients with a confirmed advanced non-small cell lung cancer (aNSCLC) diagnosis who initiated a first-line (1 L) checkpoint inhibitor (CPI) regimen in or after 2017 and had a documented PD-L1 biomarker status. Based on prior published literature, we expected an increased rwOS among CPI patients with PD-L1 positive status and/or higher PD-L1 staining than those with PD-L1 negative status or lower PD-L1 staining[17,18] in case of an unbiased estimate not affected by missingness in Hgb laboratory tests. Similarly, we further derived a second cohort of patients diagnosed with multiple myeloma (MM) who received either lenalidomide, bortezomib, and dexamethasone (VRd) or carfilzomib, lenalidomide, and dexamethasone (KRd) on first-line therapy. A clinical trial comparing these treatments found no evidence of a significant difference in rwOS or real-world progression-free survival.[19]

We applied our proposed workflow to both cohorts to assess the possibility for MCAR, MAR, or MNAR being the most likely missingness scenario. We further applied Cox proportional hazard regression models to estimate the hazard ratios (HRs) for rwOS that would result from a complete case analysis, a multiple imputation analysis using the *logreg* imputation algorithm in the mice package[14] to impute missing Hgb laboratory values and an NARFCS sensitivity analysis to illustrate potential tipping points that would lead to a rejection of the results of our primary analysis. More details on the exemplary case studies are summarized in Table S1.

RESULTS

Simulation study results

Diagnostic results

The results evaluating the AUC diagnostic are displayed in Table 1. We observed that MCAR missingness results in AUC values close to 0.5, which indicates a classifier that performs as well as random guessing. This is to be expected because under MCAR, there is no systematic difference between patients who have a missing laboratory test versus those who have one observed. Therefore, there is no relationship between the classifier features X and the missingness indicator M_{lab}. Under the other missingness mechanism, we observed AUC values meaningfully greater than 0.5, because the missingness probability is a function of observed variables. Both MAR mechanisms showed higher AUCs (0.90 and 0.92) than the other mechanisms. Because the MNAR mechanisms are partially driven by unobserved variables (either U or *lab*) the predictive power of observed variables for M_{lab} was attenuated somewhat.

Based on these results, we concluded that an AUC greater than 0.5 is a necessary condition for MAR, although not sufficient, as this can occur given MNAR data too. However, MAR was consistent with higher AUC values than those seen under MNAR scenarios. This was also previously observed in work by Beaulieu-Jones et al.[12] In practice, we recommend that this diagnostic is combined with domain knowledge of the data collection process and expected distributions in order to characterize missingness. For example, given an observed distribution of laboratory values that has fewer extreme values than expected in standard clinical care, we may conclude that the missingness mechanism is MNAR-value, even with an observed AUC diagnostic greater than 0.5.

TABLE 1 Little's test and AUC diagnostic results by simulated missingness mechanism. Little's test results are reported as probability of significant test (p value <0.05) rejecting MCAR, while AUC diagnostics are reported as mean with 95% confidence interval.

Mechanism	Little's test	AUC diagnostic
MCAR	0.03	0.49 [0.49, 0.50]
MAR v1	1	0.90 [0.90, 0.90]
MAR v2	1	0.92 [0.91, 0.92]
MNAR-unmeasured	1	0.84 [0.84, 0.84]
MNAR-value	1	0.87 [0.87, 0.87]

Abbreviations: AUC, area under the receiver operating characteristic curve; MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

Analysis results

The L1 bias results by missingness mechanism are displayed in Figure 3. Bias in regression coefficient estimates is generally reduced by applying multiple imputation, compared to complete case analysis. The exception is with MAR v1, where complete case analysis showed slightly lower bias. MNAR-value missingness resulted in the greatest L1 bias under both analytic methods. We reported the absolute bias for each variable separately in Figure 3. Here, we see that multiple imputation estimated the treatment hazard ratio unbiasedly, uniformly across missingness mechanisms. Bias was also reduced for estimating the HR of the covariate Z, but increased for the *lab* HR under MAR mechanisms. This is similarly borne out in the empirical coverage probabilities displayed in Figure 3. Under MAR, the confidence interval coverage for the *lab* HR was low under multiple imputation, whereas valid for complete case analysis. Conversely, the coverage for the *trt* HR under multiple imputation was valid across all mechanisms, including MNAR. Therefore, when a confounding variable is subject to missingness, multiple imputation can provide unbiased estimation and valid statistical inference for the treatment effect of interest.

Application to real-world cohort

Descriptive statistics

After applying the cohort-specific inclusion and exclusion criteria, 1930 and 3966 patients remained in the aNSCLC and MM cohorts, respectively. An Hgb measurement was observed for 91.7% of patients in the aNSCLC cohort and for 83% in the MM cohort. The distributions of the observed laboratory measurements are illustrated in Figure S1.

Workflow diagnostics

The results of Little's/Hotelling's test are displayed in Table 2. Overall, all p values suggested a highly significant difference in characteristics between patients with and without an observed Hgb laboratory test indicating strong evidence against MCAR. This was supported by descriptive statistics with acid sphingomyelinase deficiencies (ASMDs) greater than 0.1 for the majority of patient characteristics (Figure S2).

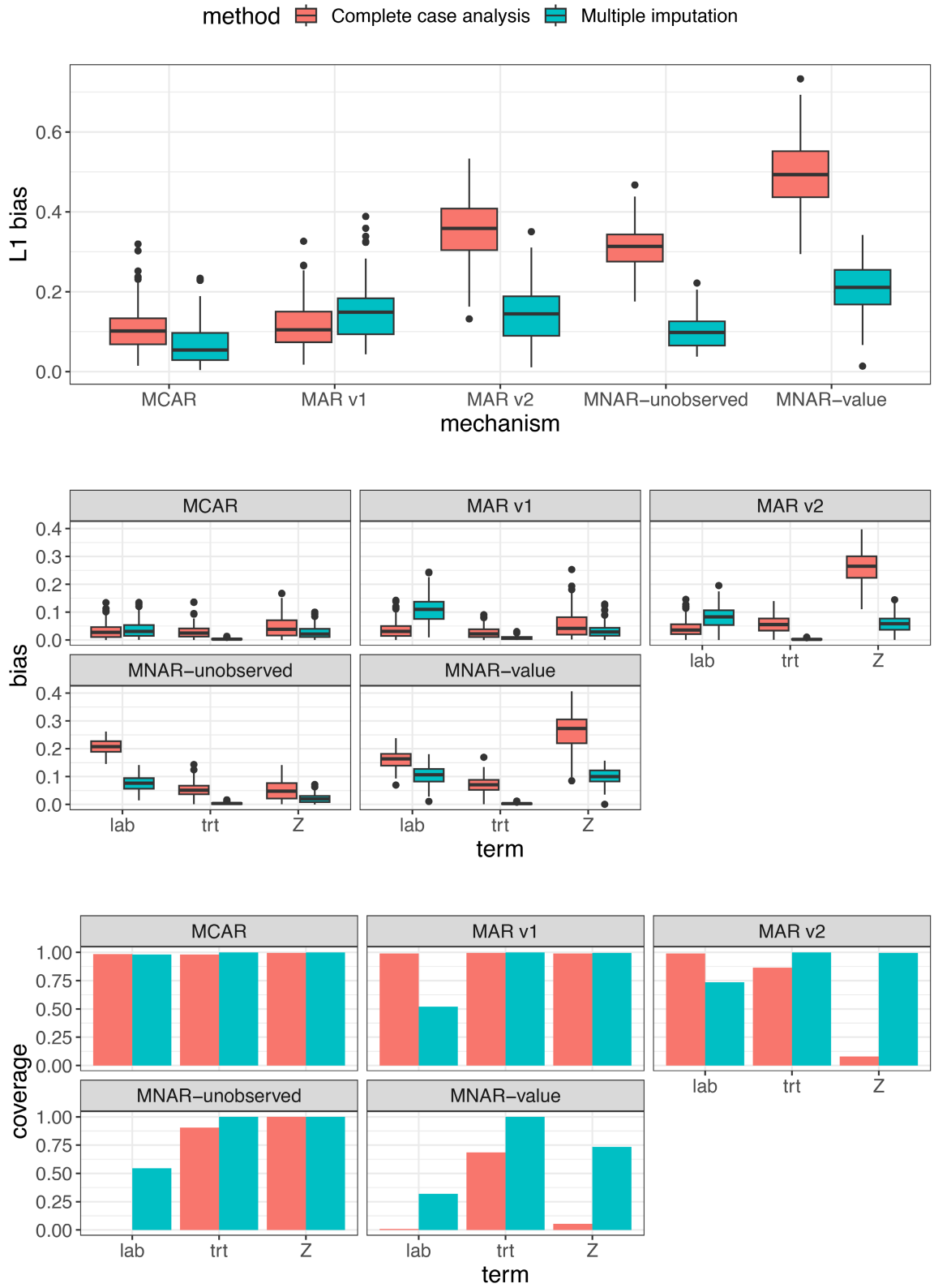

FIGURE 3 Top: Distributions of L1 bias (sum of absolute biases across all hazard ratios estimated) in simulation studies, by missingness mechanism. Middle: Distributions of absolute bias for each hazard ratio estimated (laboratory, treatment, and covariate) in simulation studies, by missingness mechanism. Bottom: Coverage probabilities of 95% confidence intervals for each hazard ratio estimated (laboratory, treatment, and covariate) in simulation studies, by missingness mechanism. MAR, missing at random; MCAR, missing completely at random; MNAR, missing not at random.

To investigate whether the missingness of Hgb may follow the MAR mechanism, we fit a classifier based on a RF model. The optimal hyperparameters in the final model were determined via grid search with five-fold cross validation (Figure S3). The resulting AUC values with corresponding 95% confidence intervals (CIs) and receiver operating characteristic curves are displayed in Figure 4. Overall, AUC values were 0.58 (aNSCLC) and 0.68 (MM), respectively, indicating some evidence for MAR laboratory values, particularly in the MM cohort. The variable importance and hence the strength of association between an observed patient characteristic and the missingness of a laboratory test was additionally determined by the mean decrease in accuracy of the prediction after iteratively removing patient characteristics one by one. The results (Figure S4) revealed a rather heterogeneous pattern suggesting that for each laboratory test and cancer type, different characteristics may have been important for the respective prediction.

TABLE 2 Results of Hotelling's test.

Cohort	Test statistic	*p* value
aNSCLC	146.29	<0.0001
MM	465.79	<0.0001

Note: Covariates compared for aNSCLC: age at index date, gender, index year, histology, group stage, smoking status, birth year, race/ethnicity, region, ECOG, age at diagnosis, age at advanced Dx, time initial Dx to index, time advanced Dx to index date, time from index date to end of follow-up, censoring indicator, PD-L1 status.Covariates compared for MM: age at index date (1 L treatment start), ECOG at index date, gender, ISS stage, index year (calendar year of 1 L treatment start), practice type, race/ethnicity, region, time diagnosis to index date, time from index date to end of follow-up, censoring indicator, line of therapy.

Abbreviations: aNSCLC, advanced non-small cell lung cancer; Dx, diagnosis; ECOG, Eastern Cooperative Oncology Group; ISS, International Staging System; MM, multiple myeloma.

Case studies analytical results and NARFCS sensitivity analysis

The results of both case studies are summarized in Table 3 and NARFCS sensitivity analyses in Figure 5. In the aNSCLC case study, investigating PD-L1 positive versus negative status, estimates derived from complete case analysis and multiple imputation yielded similar results, both suggesting an increased rwOS among patients with positive PD-L1 status (complete case HR = 0.81, 95% CI [0.68, 0.95]; multiple imputation HR = 0.79, 95% CI [0.67, 0.92]). Applying the NARFCS sensitivity analysis and imputing

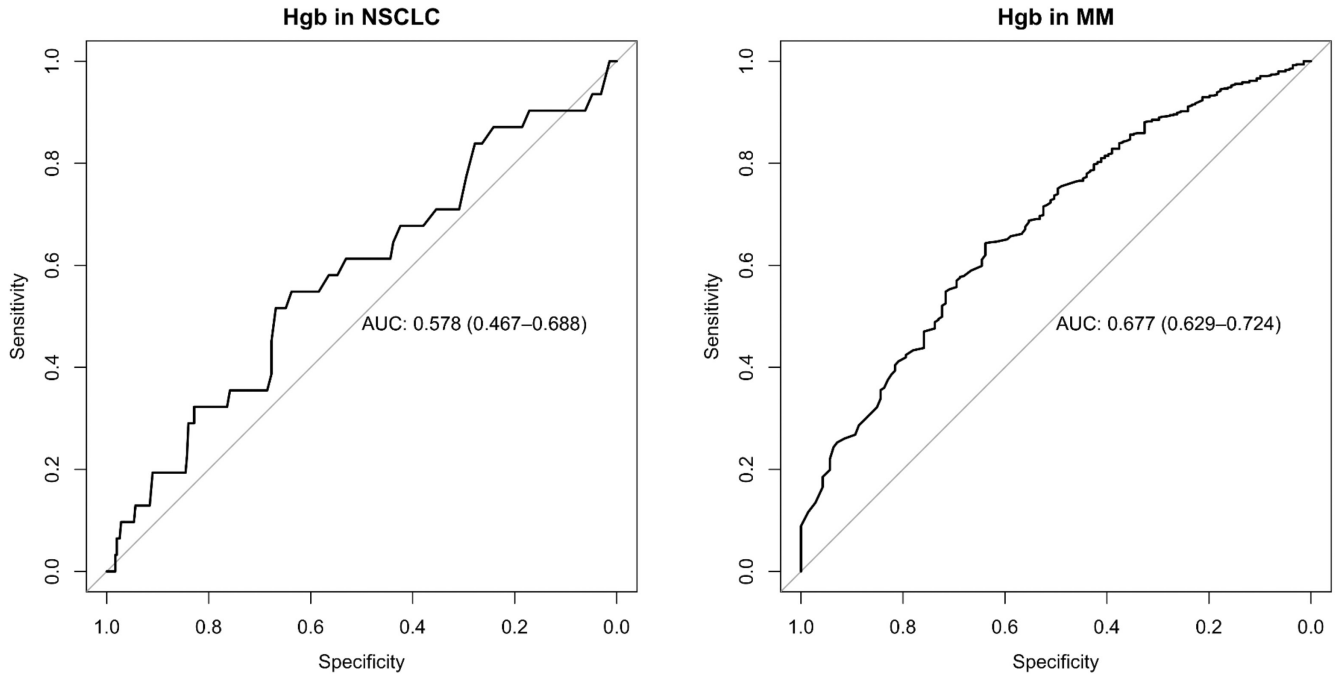

FIGURE 4 AUC diagnostic results on real-world cohorts. Covariates used in aNSCLC cohort: Age at index date, gender, index year, histology, group stage, smoking status, birth year, race/ethnicity, region, ECOG, age at diagnosis, age at advanced diagnosis, time from initial diagnosis to index, time from advanced diagnosis to index date, time from index date to end of follow-up, censoring indicator, PD-L1 status. Covariates used in MM cohort: Age at index date (1 L treatment start), ECOG at index date, gender, ISS stage, index year (calendar year of 1 L treatment start), practice type, race/ethnicity, region, time from diagnosis to index date, time from index date to end of follow-up, censoring indicator, line of therapy. aNSCLC, advanced non-small cell lung cancer; AUC, area under the receiver operating characteristic curve; ECOG, Eastern Cooperative Oncology Group; ISS, International Staging System; MM, multiple myeloma.

Cohort	Expected effect estimate (HR)	HR (95% CI)	
		Complete case analysis	Multiple imputation
aNSCLC	HR < 1.0	0.81 (0.68, 0.95)	0.79 (0.67, 0.92)
MM	HR ~ 1.0	0.98 (0.63, 1.51)	1.25 (0.87, 1.78)

Abbreviations: aNSCLC, advanced non-small cell lung cancer; CI, confidence interval; MM, multiple myeloma, HR, hazard ratio.

TABLE 3 Results of case studies investigating PD-L1 positive versus negative biomarker status among patients with aNSCLC with a first-line (1 L) checkpoint inhibitor regimen and RVD (lenalidomide, bortezomib, and dexamethasone) versus KRd (carfilzomib, lenalidomide, and dexamethasone) exposure in 1 L patients with MM.

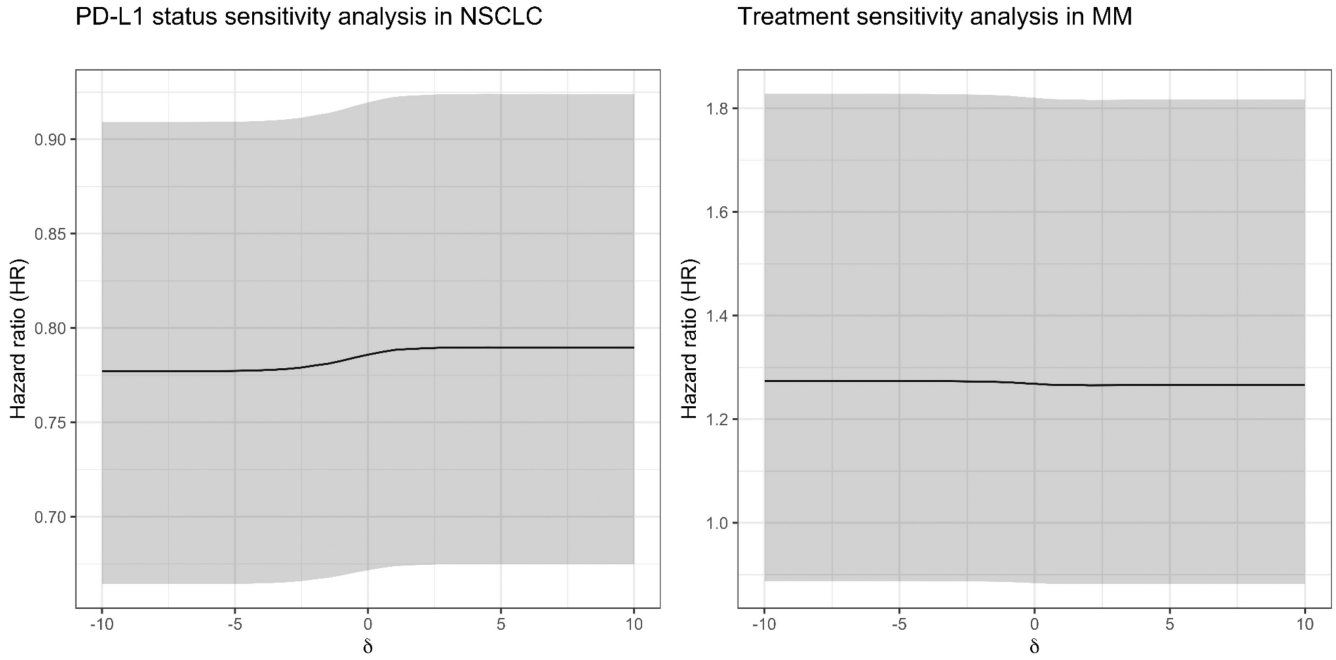

FIGURE 5 NARFCS sensitivity analysis results for real world analyses in NSCLC and MM. $\delta = 0$ indicates HR estimated under multiple imputation without any sensitivity adjustment. $\delta > 0$ indicates a shift in the imputation model where missing laboratory values are more likely to be normal than those observed; $\delta < 0$ indicates a shift in the imputation model where missing laboratory values are more likely to be abnormal than those observed. *Y*-axis displays estimated HR with 95% confidence interval. HR, hazard ratio; MM, multiple myeloma; NARFCS, not at random fully conditional specification; NSCLC, non-small cell lung cancer.

Hgb over a range of delta parameters from −10 to 10 did also not meaningfully alter this association. Therefore, even without a strong association between laboratory test missingness and observed variables, we conclude that the imputation analysis is robust to the MNAR mechanisms induced under NARFCS.

In the MM case study, directionally different results were obtained for the complete case (HR for KRd = 0.98, 95% CI [0.63, 1.51]) and multiple imputation analyses (HR for KRd = 1.25, 95% CI [0.87, 1.78]). However, the lack of statistical evidence for a significant association was concordant. As with the NSCLC case study, applying the NARFCS sensitivity analysis over a range of sensitivity parameters did not alter the estimated HR compared to the result obtained by multiple imputation alone. Given this robustness, and strong evidence against MCAR, we have confidence in using the imputation analysis results.

DISCUSSION

We developed a systematic workflow for diagnosing missing data mechanisms, which we validated in a de novo simulation study and applied to aNSCLC and MM cohorts derived from a real-world de-identified EHR-derived database. This workflow can be used to inform the appropriateness of complete-case, imputation, and sensitivity analysis approaches. To our knowledge this is the first study to investigate two different MNAR scenarios with different missingness assumptions, leading to differences in diagnostic results and bias. Moreover, our simulations showed unbiased results under multiple imputation for the treatment HR, even under MNAR mechanisms. This is consistent with previous work showing that the bias resulting from omitting confounders with only moderate prevalence and confounding strength was negligible.[20] Furthermore, very few studies have included the NARFCS

algorithm as an additional component to strengthen their conclusions. All methods can be implemented using standard software.[14]

This work tackles challenges caused by heterogeneous data capture that can result in bias and greater uncertainty in analyzing RWD. For the use of real-world evidence (RWE) in the regulatory realm, both the US Food and Drug Administration and the European Medicines Agency have pointed out the challenges posed by missing data, recommending a series of approaches based on methods such as inverse probability weighting or likelihood-based methods[21] as well as multiple imputation approaches,[22] given full transparency regarding the assumptions required for validity of these methods,[23-28] and accompanied by sensitivity analyses.[25] Recent literature has noted that few studies address how missingness assumptions are actually checked, or how the missingness is handled in analyses; studies that propose recommendations do so in a non-systematic fashion.[12,19,29,30,31,32,33] Our workflow is an easy-to-implement approach to check analytical assumptions around the potential mechanisms of missingness of critical variables before conducting further downstream analyses.

The AUC-based diagnostic for MAR can be implemented with any classifier and is easy to interpret as a measure of association between missingness and observed variables. This is borne out in our simulations, where missingness under MAR and MNAR mechanisms was predictable. MAR consistently achieved higher AUC diagnostics than the MNAR mechanisms, which is in line with previous work.[12] These two missingness mechanisms are not formally distinguishable from observed data alone, so we recommend that domain knowledge of the data-generating process is also taken into consideration to assess the plausibility of MNAR. NARFCS sensitivity analysis can also be applied in more ambiguous situations to rule out meaningful departures of the downstream analytical results under a possible MNAR mechanism. In our case study examples, for instance, a potential MNAR mechanism would have not meaningfully changed the estimates of our main analyses, both of which were in line with results reported in the literature and coming from clinical trials.

A limitation of this study is that we focused on the missingness of one laboratory test at a time and assumed that other variables were fully observed. In reality, many variables in a dataset can be missing under different missingness mechanisms. Nevertheless, the proposed workflow may still diagnose the potential missingness mechanisms of the most critical variables for a given research question and the NARFCS sensitivity analysis can be extended to multiple variables.[13]

Possible future extensions of this study may focus on multivariable scenarios and explicit nonlinear and non-additive variable relationships. In addition, the influence of various differential and non-differential missingness scenarios may be of particular interest in the context of modeling and evaluating treatment effects in RWE studies. Finally, extensions of our investigations to observational causal inference analyses may be of high interest for potential future research.

AUTHOR CONTRIBUTIONS

Wrote Manuscript: A.S. and J.W. Designed Research: All authors contributed. Performed Research: A.S., J.W., P.Y., and S.C. Analyzed Data: A.S. and J.W. Contributed New Reagents/Analytical Tools: N/A.

ACKNOWLEDGMENTS

The authors acknowledge the assistance of Hannah Gilham of Flatiron Health for providing writing and editing support for this paper.

FUNDING INFORMATION

This work was supported by Flatiron Health, which is an independent subsidiary of the Roche Group.

CONFLICT OF INTEREST STATEMENT

A.S., P.Y., C.J., M.S., and S.C. all report employment in Flatiron Health Inc., which is an independent subsidiary of the Roche Group, and stock ownership in Roche. J.W. reports employment at Hoffmann-La Roche, and stock ownership in Roche. M.T. reports employment at Genentech, a Member of the Roche Group, and stock ownership in Roche.

REFERENCES

1. Miksad RA, Abernethy AP. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin Pharmacol Ther*. 2018;103:202-205.
2. Becker T, Weberpals J, Jegg AM, et al. An enhanced prognostic score for overall survival of patients with cancer derived from a large real-world cohort. *Ann Oncol*. 2020;31:1561-1568.
3. Sv B. *Flexible imputation of missing data*. 2nd ed. CRC Press; 2018.
4. Carpenter JR, Smuk M. Missing data: a statistical framework for practice. *Biom J*. 2021;63:915-947.
5. Lee KJ, Tilling KM, Cornish RP, et al. Framework for the treatment and reporting of missing data in observational studies: the treatment and reporting of missing data in observational studies framework. *J Clin Epidemiol*. 2021;134:79-88.
6. Cornish RP, Macleod J, Carpenter JR, Tilling K. Multiple imputation using linked proxy outcome data resulted in important bias reduction and efficiency gains: a simulation study. *Emerg Themes Epidemiol*. 2017;14:14.
7. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157-166.
8. Schober P, Vetter TR. Correct baseline comparisons in a randomized trial. *Anesth Analg*. 2019;129:639.

9. Little RJA. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*. 1988;83:1198-1202.
10. Hotelling H. The generalization of Student's ratio. *Ann Math Stat*. 1931;2:360-378.
11. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol*. 2006;59:1087-1091.
12. Beaulieu-Jones BK, Lavage DR, Snyder JW, Moore JH, Pendergrass SA, Bauer CR. Characterizing and managing missing structured data in electronic health records: data analysis. *JMIR Med Inform*. 2018;6:e11.
13. Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, White IR. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Stat Med*. 2018;37:2338-2353.
14. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Soft*. 2011;45:1.
15. Birnbaum B, Nussbaum N, Seidl-Rathkopf K, et al. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research *arXiv*: Preprint posted online January 13. 2020.
16. Ma X, Long L, Moon S, Adamson BJS, Baxi SS. Comparison of population characteristics in real-world clinical oncology databases in the US: flatiron health, SEER, and NPCR. *medRxiv*: Preprint posted online May 30, 2020. 2020.
17. Khozin S, Miksad RA, Adami J, et al. Real-world progression, treatment, and survival outcomes during rapid adoption of immunotherapy for advanced non-small cell lung cancer. *Cancer*. 2019;125:4019-4032.
18. Shen X, Zhao B. Efficacy of PD-1 or PD-L1 inhibitors and PD-L1 expression status in cancer: meta-analysis. *BMJ*. 2018;362: k3529.
19. Kumar SK, Jacobus SJ, Cohen AD, et al. Carfilzomib or bortezomib in combination with lenalidomide and dexamethasone for patients with newly diagnosed multiple myeloma without intention for immediate autologous stem-cell transplantation (ENDURANCE): a multicentre, open-label, phase 3, randomised, controlled trial. *Lancet Oncol*. 2020;21:1317-1330.
20. Nguyen T, Collins GS, Spence J, et al. Magnitude and direction of missing confounders had different consequences on treatment effect estimation in propensity score analysis. *J Clin Epidemiol*. 2017;87:87-97.
21. US Food and Drug Administration. *Framework for FDA's real-world evidence program*. US Food and Drug Administration; 2018.
22. National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. 2010. Accessed April 17, 2023. https://www.ncbi.nlm.nih.gov/books/NBK209904/
23. US Food and Drug Administration. *Adjusting for covariates in randomized clinical trials for drugs and biological products guidance for industry [Draft Guidance]*. 2021.
24. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. *N Engl J Med*. 2012;367:1355-1360.
25. US Food and Drug Administration. *Center for drug evaluation and research, (CDER). Review of 212018Orig1s000*. Accessed April 17, 2023. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/212018Orig1s000MultidisciplineR.pdf
26. US Food and Drug Administration. *Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims*. 2009.
27. Alcini P, Candore G, Lehmann M, et al. *The HMA/EMA task force on big data: data analytics subgroup report*. 2019. Accessed April 17, 2023. https://www.ema.europa.eu/en/documents/minutes/hma/ema-joint-task-force-big-data-summary-report_en.pdf
28. Kohl S. Joint HMA/EMA task force on big data established. *Eur J Hosp Pharm*. 2017;24:180-190.
29. Faria R, Wailoo A, Manca A, Alava MH. *NICE DSU Technical support Document 17: the use of observational data to inform estimates of treatment effectiveness in technology appraisal: methods for comparative individual patient data*. Accessed April 17, 2023. https://www.sheffield.ac.uk/sites/default/files/2022-02/TSD17-DSU-Observational-data-FINAL.pdf
30. Hunt NB, Gardarsdottir H, Bazelier MT, Klungel OH, Pajouheshnia R. A systematic review of how missing data are handled and reported in multi-database pharmacoepidemiologic studies. *Pharmacoepidemiol Drug Saf*. 2021;30:819-826.
31. Carroll OU, Morris TP, Keogh RH. How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. *BMC Med Res Methodol*. 2020;20:134-137.
32. Tierney NJ, Harden FA, Harden MJ, Mengersen KL. Using decision trees to understand structure in missing data. *BMJ Open*. 2015;5:e007450.
33. Baron JM, Paranjape K, Love T, Sharma V, Heaney D, Prime M. Development of a "meta-model" to address missing data, predict patient-specific cancer survival and provide a foundation for clinical decision support. *J Am Med Inform Assoc*. 2021;28:605-615.
34. Mohan K, Pearl J, Tian J. Graphical models for inference with missing data. *Adv Neural Inf Process Syst*. 2022; 26.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sondhi A, Weberpals J, Yerram P, et al. A systematic approach towards missing lab data in electronic health records: A case study in non-small cell lung cancer and multiple myeloma. *CPT Pharmacometrics Syst Pharmacol*. 2023;12:1201-1212. doi:10.1002/psp4.12998